Geoff Huston
November 2017

RIPE 75

RIPE held its 75th meeting in Dubai in mid-October. As usual there was a diverse set of presentations covering a broad range of activities that are taking place on today's Internet. The topics include issues relating to network operations, regulatory policies, peering and interconnection, communications practices within data centres, IPv6, the DNS, routing and network measurement. If that's not enough, the topic of the Internet of Things has been added as a Working Group in the RIPE pantheon. If you add address policy, database and RIPE services to the mix you get a pretty packed five days with topics that would appeal to most Internet folk

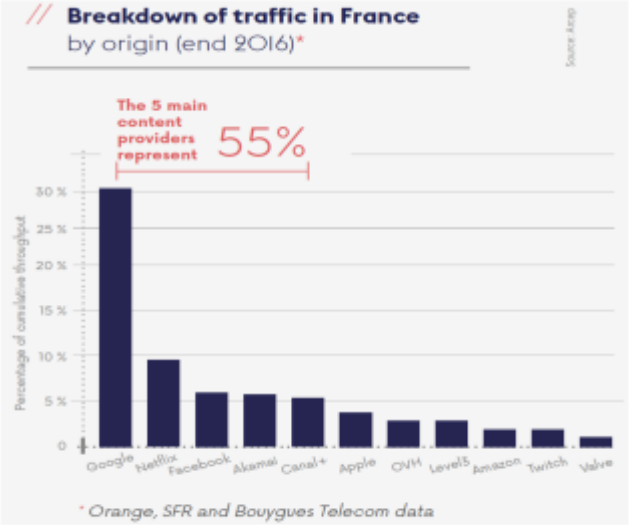
Here's my impressions of the meeting, covering just a small subset of the presentations made at the meeting.

Measurement Tools: Richard Cziva presented on his research work in building a real time TCP latency measurement tool capable of visualising traffic carried on a 10G circuit. This was a fascinating triumph of using open source tools combined with some neat hacking. This "RURU" tool starts with a passive optical fibre tap on a 10G service. This is fed into a server equipped with an Intel 10Gps network card. This is a critical part of the exercise, as the potential packet rate rate easily exceeds the processing capacity on the server. However, the tool uses the Intel processing card and the DPDK library to perform the packet processing. The processing is relatively straightforward, in that it is looking for TCP handshakes, and recording the times between the SYN and the SYN/ACK and the SYN/ACK and the ACK. The challenge is to perform these measurements and then feed them into a visualisation display system in real time so that the entire profile of connections that take place on a high speed trunk fibre circuit can be visualised as they occur. I'm not entirely sure of its utility as a network management tool, but the tool's graphics were indeed very cool!

RURU Screen Shot - Richard Cziva

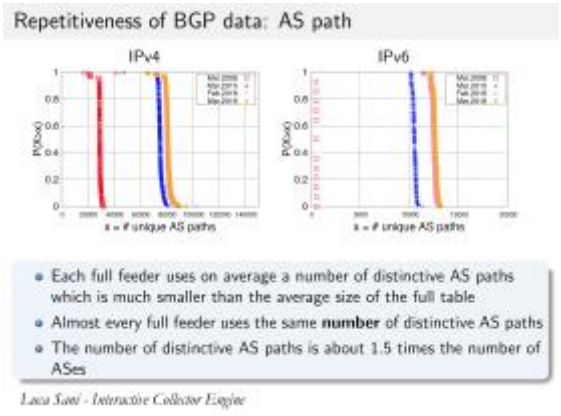
Regulation: ARCEP reported on a couple of recent issues concerning interconnection disputes in France. The first concerned a dispute between the transit provider Cogent and Orange, a major access provider in France and elsewhere, but I understand that only the French Orange was involved in this particular dispute. The terms of the peering agreement stipulated "approximately equal" traffic volumes to be exchanged between the two parties, but when Cogent carried traffic from Megaupload, the traffic balance shifted significantly and Orange refused to provide capacity increases to alleviate congestion on the peering link. The French Competition Authority rejected all of Cogent's allegations, finding that requesting payment in the case of a highly asymmetric traffic exchange does not in itself constitute an

anti-competitive practice. However, the French Competition Authority had accepted a transparency commitment subscribed to by Orange in order to clarify the relationship between Orange's domestic network and its transit activity. The dispute with Cogent started in 2011 and lingered on until the Paris Court of Appeal rejected Cogent's claims, effectively confirming the findings of the Competition Authority. A similar dispute arose in 2012 between Google and Free, a major retail ISP in France. In this case Google complained over the alleged use of traffic shapers by Free on YouTube streaming traffic. Free is alleged to have limited YouTube traffic, and in early January 2013, it installed ad-blocking software on its routers, something it later ended after regulatory intervention. Orange, Cogent, Telia, NTT Google and Free have all had disputes over forms of interconnection in France and the regulator is favouring a role of facilitating case-by-case dispute resolution rather than either using industry-wide regulatory edicts or simply referring the parties into the legal system. The regulator does not see itself as shaping the market, but encourages the various service providers in each case to find an acceptable outcome that is market-based. In my view, this position does seem to create a somewhat awkward relationship with the French ISP industry, and doubtless the three-way tensions between transit providers, access networks and content networks will continue. To underline the importance of content in this market, the report noted that some 70% of the traffic in the French IP networks was streaming traffic. Most of this traffic originated in the large Content providers and the CDN aggregators, so Google dominates with 55% share of the overall content volume, followed by the usual mix of Netflix, Facebook, Akamai, Apple and Amazon, together with the regional OVH and the French Canal+



Samih Souissi, Arcep, "State of the Internet in France: Data Interconnection Market Evolution"

Routing and Measurement: The Route Views project has been in place for many years, and it is a very valuable resource relating to the behaviour of BGP over the years. This was complemented by the RIPE NCC's Routing Information Service, and more recently PCH has also started a project of operating route collectors at exchanges. Isolario is another route collector project, operated by IIT-CNR in Italy. Its distinctive feature is the ability to perform near-real time feeds out of the route collector, using a path attribute compression technique to allow the system to operate efficiently and quickly. I have always use the route collector data to ask questions about BGP events in the past. The Isolario system allows this to extend to BGP events happening in close to real time. One of the features that this approach exploits is the observation that the new information load in BGP is far less than the traffic load. Much of what BGP provides in updates is not unique to a simple prefix, and not unique over time. I'm not sure how I could use such a real time feed myself, but I'm keen to give it a try!



Luca Sani - Isolario Collector Engine

BGP. The Border Gateway Protocol is a venerable protocol these days, and has been holding the Internet together for some 25 years. At its heart it's the same protocol, but it has had its fair share of adornment with sundry bells and whistles over the years. Nokia's Greg Hankins reported on some recent work in refining aspects of the way that BGP works. I must admit to a certain level of fascination with the BGP protocol, so I'll take a little time here to work through the finer details of his presentation

- RFC7999 defines a "remote Blackhole" community. An operator may attach this community to a prefix that is the target of an incoming traffic attack. The effect of this is that neighbouring networks that honour this attribute will not pass traffic towards the originating AS, but instead will locally sink the traffic. This means that while the target prefix is still offline, collateral damage that may have overwhelmed all available capacity leading to the target network will be mitigated in this approach. A somewhat perverse side effect is that this has been cited as a reason for deaggregation of route advertisements, as black holing the more specific route will have less of a side effect on services in adjacent addresses than if the aggregate were to be blackholes.

- The next recent change is RFC8192, which changes the default neighbour configuration to an implicit DENY ALL policy clause. This way, route advertisements will not occur until explicit route policy clauses are added to the router config. It's a bit like setting the switches to OFF before applying power to a circuit.

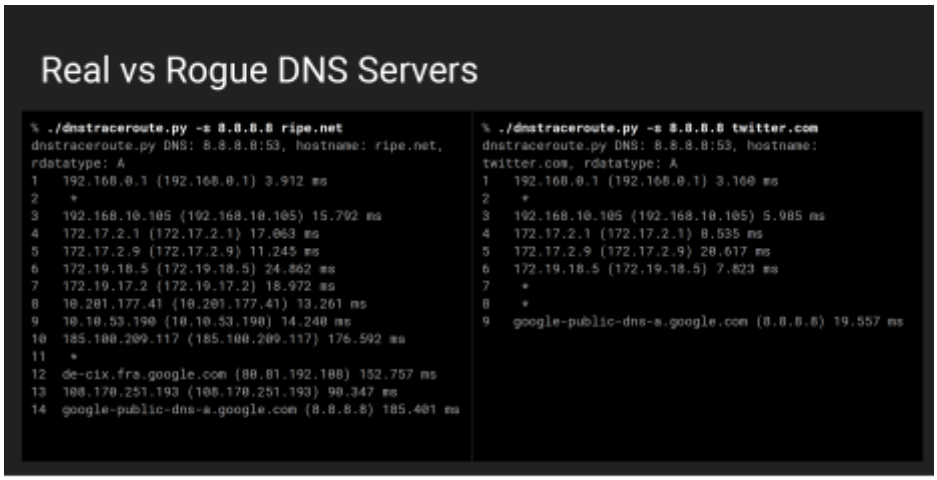
- RFC8092 cuts through years of pointless IETF quibbles over exactly how we would integrate 32 bit AS numbers into BGP communities by defining a small set of community formats that mimic the functionality provided by 16t bit communities. Who would've thought that it would take so long to adopt the obvious approach!

- RFC8203 does not appear to be all that useful at first glance. When you shut down a BGP session you are expected to prove some readable text as to why the session is being closed, and this will be transmitted to the other side of the peering session. But there is the potential for operators to use standard message templates to show that a planned event is taking place and the session is anticipated to be restarted at a certain time and integrate this into network management systems.

- draft-ietf-grow-bgp-gshut proposes form of "graceful" shutdown where an attribute can be attached to an advertised prefix that signals to its peers to drop the local preference of an advertised prefix. This should trigger the peer to select an alternate route for the prefix if one exists, so that a subsequent withdrawal of the prefix would have minimal impact on all peers.

- draft-ietf-grow-bgp-session-culling proposes some form of managed session shutdown such that the control plane traffic is suspended and a hold timer is started. When this hold timer expires, the traffic over the link is also suspended. Between the conventional operation of BGP keepalives and grateful shutdown, this looks to me to be a largely superfluous adornment to BGP and the fact that the BGP world appears to have got on just fine for 30 years without it does not make a convincing case in my mind that this is a necessary addition to BGP.

DNS: Babak Farrokhi from Iran gave a fascinating presentation on the DNS in a curious case of broken DNS responses. His investigations started with failed outgoing mail, and posing queries to both his local DNS resolver and to the IP address of Google's Public DNS services also resulted in incorrect responses. In one case the response was a badly formatted DNS



Barbak Farrokhi - A curious case of broken DNS responses

packet and the time difference between the query and the response was unusually short. From this came the first rather clever tool: dnsping. It looks like ping but rather than sending UDP packets or ICMP echo requests, dnsping sends DNS queries and times how long to get the matching response. Because of DNS caching repeated queries to the same recursive resolver with the same query name should produce elapsed times comparable to that of a simple ping to the same IP address. When in this case the dnsping shows substantially smaller response times than ping for the same resolver IP address then there is a good case to suspect some form of DNS interception is being used in the network. But where? This lead to the next tool he wrote: dnstraceroute. It's the same principle as conventional tracerout: namely sending packets within increasingly longer TTL values in the IP header and collecting the ICMP "time exceeded" responses that come back. This creates a hop-by-hop packet trace from the test point right to the interception location! That's very cunning! And of course there is also the domain "maxmind.test-ipv6.com". When you query this domain for a TXT record the response is the IP address, country and AS number of the resolver which performed the query on your behalf. This provided the platform for a test using the RIPE Atlas probes: What proportion of DNS queries to Google's Public DNS Server are being redirected to elsewhere? The answer is somewhat disturbing, in that across 484 probes using UDP and IPv4 some 2% of the queries (9 in total) were being redirected. The number falls to 1% (5 in total) when using TCP. What can you do about this? If your concern is that your local DNS environment is being compromised, or if you are concerned that your local DNS environment is leaking information about your queries then perhaps the best answer is to use a local DNS resolver that supports DNS privacy. This will place your queries in an encrypted session with a remote DNS recursive resolver, relying on encryption to effectively hide your queries from local inspection or tampering. The DNS Privacy Project (https://dnsprivacy.org) is a good place to start, and for a local DNS resolver that supports this approach, try Stubby (https://dnsprivacy.org/wiki/display/DP/DNS+Privacy+Daemon+-+Stubby). As Babak observed: "Don't trust your upstream, encrypt as much as possible."

Transport and Encryption: On this subject of "encrypt as much as possible" was a related lightning talk on QUIC. QUIC is a protocol originally developed by Google and rolled out in the Chrome browser since 2014. It evidently is now 35% of all Google traffic and 7% of all Internet traffic. Among many attributes, QUIC has a minimal exposure of information. It is part of the observation about the pervasive use of intercepting middleware on the Internet and the regressive stance taken by IP carriers. Every exposed bit in a transport session is a bit that will be exploited by middleware and will be interested in a fixed manner, so its a bit that can never change thereafter. Equally every exposed bit can be used against us in the future. So Quic's stance is to expose as little as possible. Obviously this approach has generated some reaction. Network operators are, predictably, unhappy about this development, and there is some resistance in the IETF to standardise this universal approach to encryption of the entire session. My opinion? I'm with Google on this. The only defence against the ubiquitous wiretapping mentality of IP carriers is to encrypt it all. So lets push everything below the visibility of this pernicious middleware!

Building Networks: Its been many years since small teams in the academic and research environment built Internet networks from scratch. But thats exactly what SUNET has done in Sweden, and we heard of the work in Sweden to install a green fields 100G backbone network for the national academic and research sector. In their case a number of existing arrangements all ended at much the same time, including IRUs on fibre lines, write off of the book value of existing equipment, DWDM leases and such. Given this it was possible to design a new network without legacy elements for the country. The 8.2K fibre network uses Ramen optical amplifiers with care attention to loss and reflection. In the high speed world reflection is a very significant concern. In the line cards for the electronics the line card itself contains the optical transport interface and DWDM. The allows for fewer components in the network. The basic network design is a ladder with cross-connects, ensuring diversity in the ladder paths. As with many networks these days some 90% of the configuration in the routers is centrally orchestrated rather than manually configured.

Encryption and Privacy: Integrity and privacy in the DNS continue to be a concern, and NLnet Labs' Benno Overeinder looked at the current state of DNS resolvers in attempting to address these issues.

The first is that of privacy of DNS requests, addressing concerns that a user's DNS queries are being inspected and potentially intercepted. The best response we have in this space is to use DNS over TLS, or, potentiality DTLS (DNS over TLS adapted for UDP), assuming that DTLS implementations are available for use at some point. The crypto is much the same, is that there is a TLS handshake where the client needs to authenticate the server's credentials against a local trust anchor store. The problem is that with so many CAs issuing certificates, the implicit assumption that all CAs never lie has proved to be a remarkably poor assumption. In this scenario, DANE is a decent response, where the credentials of the named server are contained in the DNS, and are themselves protected by DNSSEC. This creates an obvious circularity of using the DNS to check the credentials of a DNS server and the potential of being led astray is obvious. The response is to make use of DNSSEC chain extensions in the TLS exchange (https://tools.ietf.org/html/draft-ietf-tls-dnssec-chain-extension) and rely on the DNS Trust Anchor alone to validate the presented credentials. This approach is useful in the context of using a trusted recursive resolver and setting up a secure and private communication channel with that resolver.

This is a small selection of the material presented at RIPE 75. The complete set of presentations can be found at: https://ripe75.ripe.net

Author

Geoff Huston B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.

www.potaroo.net

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.